International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Deepfake Detection

[1]FATHIMA ROSNA T A, [2]FATHIMA HANNATH, [3]MEGHA K S, [4]HARIPRIYA, [5]Ms SREEJI

[1]Student, [2]Student, [3]Student, [4]Student,[5]Assistant Professor (CSE)
Computer Science and Engineering Department,
Nehru College of Engineering and Research Centre (NCERC), Thrissur, India

---

Abstract - In the age of AI, deepfake technology has emerged as a critical challenge, producing hyper-realistic yet manipulated images, videos, and audio that blur the line between reality and fabrication. These deepfakes pose serious risks, including spreading misinformation, compromising privacy, and undermining trust in digital media. Our project, DeepShield, is a robust web based platform developed to detect and mitigate the impact of deepfake content. Designed for accuracy and efficiency, DeepShield identifies and classifies manipulated media across images, videos, audio files, and social media links. Leveraging advanced AI and machine learning, the system examines key indicators such as facial features, audio-visual inconsistencies, and digital artifacts, ensuring high detection precision. By processing large datasets rapidly, DeepShield offers a reliable tool for individuals, media professionals, and organizations to combat the spread of deceptive content. Through its user-friendly design, DeepShield provides a practical solution to the growing threat of deepfake technology, reinforcing authenticity in the digital age.

Key Words: Deepfake, Detection, Misinformation, AI (Artificial Intelligence), Privacy

1. INTRODUCTION

Deepfake technology, which utilizes AI to create manipulated media, has become a growing concern due to its ability to produce highly realistic but deceptive content. Deepfake media, including altered videos, images, and audio, poses serious risks to privacy, information integrity, and public trust. The spread of such fabricated content across social platforms has fueled misinformation, damaged reputations, and presented new challenges to digital security.

DeepShield addresses this issue by providing a web-based solution for detecting and mitigating deepfake content. The platform is designed to identify manipulated media across various formats—images, videos, audio files, and social media links—offering a comprehensive approach to deepfake detection. DeepShield employs advanced AI and machine learning techniques that allow it to recognize subtle indicators of manipulation, including inconsistencies in facial features, flaws in audio-visual synchronization, and unique digital artifacts that signal content tampering.

The system's detection process works by analyzing input media on multiple levels. For instance, DeepShield examines facial expressions, movements, and background patterns to detect visual anomalies, while audio analysis methods identify irregularities in sound and timing. By examining these layered aspects of deepfake media, DeepShield achieves high accuracy in identifying falsified content. In addition to its detection capabilities, DeepShield is built for scalability, allowing it to process large datasets efficiently. Its cross-platform functionality supports users across various digital environments, making it an accessible tool for individuals, journalists, and organizations alike. Through DeepShield, users gain a practical resource to verify content, ensuring that digital media remains secure and credible. This project aims to create a system that accurately detects deepfakes and strengthens the integrity of digital content across platforms. By combining precision detection methods with efficient data processing, DeepShield represents a proactive solution to the growing problem of deepfake technology.

2. LITERATURE REVIEW

[1] **DeepFake MNIST+: A DeepFake Facial Animation Dataset**

DeepFake MNIST+ is a dataset designed to advance deepfake detection, specifically targeting facial animation attacks. Created by Jiajun Huang, Xueyu Wang, Bo Du, Pei Du, and Chang Xu, it consists of 10,000 facial animation videos intended to challenge liveness detection systems used in secure applications like biometric-based payment verification. The dataset helps researchers develop methods to detect deepfakes that mimic realistic facial animations. A baseline detection method is also proposed for evaluating detection models.

However, the dataset has limitations: it focuses solely on facial animations and does not include identity-swapping deepfakes. Additionally, its effectiveness decreases with video compression or lower-quality footage, making models trained on this dataset less reliable for real-world deepfakes, which are often compressed on social media platforms.

[2] **A Survey on Deepfake Video Detection**

This survey, authored by Peipeng Yu, Yanpeng Cai, Zhihua Xia, Yunqing Shi, and Weiwei Sun, reviews key techniques for detecting deepfake videos, categorizing them into five approaches: general network-based methods that use neural networks to identify deepfake patterns, temporal consistency-based methods that detect inconsistencies across video frames, visual artifacts-based techniques targeting pixel-level abnormalities, camera fingerprints-based methods that look for unique characteristics left by real cameras, and biological

signals-based approaches that focus on altered biological cues like blinking or heart rate in deepfakes. The survey also discusses challenges in deepfake detection, including the need for better model generalization, interpretability, and methods that work within practical time constraints. Recent advancements like multi-task learning and anti forensic approaches show promise, but limitations such as overfitting to specific datasets, the complexity of black-box models, and high computational costs still hinder the real-time application of these methods.

[3] **DeeperForensics-1.0 : A Large-Scale Dataset for Real-World Face Forgery Detection**

DeeperForensics-1.0, developed by Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy, is a comprehensive dataset designed to tackle forgery detection challenges by simulating real-world manipulations. It utilizes a unique DF-VAE framework to generate high-quality, manipulated videos, improving the realism and applicability of detection models. The dataset has been tested with baseline detection methods, such as 3D CNNs and ResNet, offering a solid benchmark for evaluating detection performance. However, it faces limitations, including generalization issues when applied to different types of deepfakes, sensitivity to compression affecting accuracy, and high computational costs, limiting its practicality for low-resource or real-time applications.

[4] **DFT-MF: Enhanced Deepfake Detection Using Mouth Movement and Transfer Learning**

The DFT-MF method, developed by Ammar Elhassan, Mohammad AlFawa'reh, Mousa Tayseer Jafar, Mohammad Ababneh, and Shifaa Tayseer Jafar, detects deepfake videos by focusing on mouth and teeth movements during speech, which are often manipulated in deepfakes. The method uses transfer learning with pre-trained models like CNN, DenseNet, and EfficientNet to improve detection accuracy. Evaluated on datasets like Celeb-DF, Deepfake Vid-TIMIT, and UADFV, DFT-MF performs well in real-world scenarios. However, it faces limitations, including poor generalization on unseen datasets, sensitivity to compression, and a focus on identity-swapping deepfakes, making it less effective for other manipulations. Additionally, the model is vulnerable to adversarial attacks. Despite these challenges, DFT-MF offers a promising approach to deepfake detection but requires further improvements for broader applicability and robustness.

[5] **DeepInsights of Deepfake Technology**

Deepfake technology, powered by computer vision and deep learning, enables the creation of highly realistic fake videos, images, and manipulated voices. While it offers intriguing possibilities, it poses significant risks, including the spread of misinformation, manipulation, and societal harm. The technology uses advanced algorithms to produce content that's nearly indistinguishable from real media, making detection challenging. This review explores the mechanisms, impact, and implications of deepfakes, addressing questions about their creators, potential benefits, and challenges. It highlights that, despite its dangers, deepfake risks can be mitigated through strict regulations, robust detection methods, and increased awareness.

[6] **Deep Learning for Deepfakes Creation and Detection**

Deep learning has enabled the development of technologies like deepfakes, which create fake images and videos nearly indistinguishable from real ones. While offering powerful applications, deepfakes pose significant risks to privacy, democracy, and national security by spreading misinformation and manipulating public opinion. This paper surveys the algorithms used for deepfake creation and detection, discussing the challenges in identifying deepfakes, emerging research trends, and future directions. It provides an overview of current detection methods and highlights gaps that need to be addressed to develop more effective systems for combating deepfakes.

[7] **DeepFake Detection for Human Face Images and Videos**

DeepFake technology, powered by advanced deep learning, enables the creation of hyper-realistic fake images and videos by manipulating facial features, posing significant risks to privacy, security, and public trust. This paper reviews deepfake creation methods, which are categorized into five types, and the use of deep neural networks (DNNs) in detecting fake content. It discusses the role of deepfake datasets in improving both creation and detection, highlighting the challenge of developing generalized detection models that can handle diverse media and conditions. The paper emphasizes the need for robust, adaptable detection systems to keep pace with the evolving threat of deepfakes.

[8] **GAN-Based Model of Deepfake Detection in Social Media**

Deepfake technology uses *Generative Adversarial Networks (GANs)* to swap identities in videos and images, making it challenging to differentiate between real and fake content. GANs consist of a generator creating synthetic images and a discriminator evaluating them, improving the quality of generated media. While deepfakes have potential in entertainment, they pose serious risks in privacy violations, misinformation, and public trust. This paper explores the use of pre-trained GANs for deepfake creation and focuses on detecting them with *Deep Convolutional GANs (DCGANs)*, which enhance the quality and authenticity of the generated content.

3. PROBLEM STATEMENT

The rapid advancement of deepfake technology has introduced serious challenges across fields like media, entertainment, politics, and cybersecurity. Deepfakes—hyper-realistic manipulated video and audio content generated with AI—are increasingly used in harmful ways, such as spreading misinformation, committing identity theft, and facilitating harassment. As these tools evolve, they become harder to detect, posing a threat to the credibility of digital media and potentially eroding public trust. The difficulty in verifying online content heightens the urgency for effective, accessible deepfake detection solutions capable of keeping pace with this fast-developing technology. Existing detection tools, however, face significant limitations that impact their usability and effectiveness.

DeepFaceLab, for instance, is a widely used open-source software for deepfake creation and analysis. Designed primarily for face-swapping and synthetic media generation, it offers valuable insights into the creation of deepfakes but lacks functionality as a detection tool. While it serves as a powerful educational and research tool, DeepFaceLab's focus on generation over detection leaves a gap in the fight against deepfake threats. As deepfake techniques become increasingly advanced, the absence of robust detection frameworks allows for more subtle and high-quality manipulations that evade traditional methods. This gap highlights the urgent need for a more reliable, adaptable, and user-friendly approach to deepfake detection.

4. PROPOSED SYSTEM

System Design :

The proposed deepfake detection system is structured to provide a comprehensive, two part analysis of video content, using state-of-the-art machine learning models. The design incorporates a modular approach with distinct training and prediction phases, enabling efficient detection of manipulated content with high accuracy.

The system begins by loading video data, which includes both real and fake samples, from a dataset repository. The primary purpose of this stage is to preprocess the raw video data into a format suitable for training and detection tasks. Preprocessing involves several critical steps.

• **Splitting Video into Frames** : Each video is broken down into individual frames to facilitate frame-level analysis. This granular approach allows for detailed examination of each moment within the video.

• **Face Detection and Cropping** : Using a face detection model, the system isolates the face region within each frame. This step is essential because facial features are typically the most affected in deepfake manipulation, and focusing on these regions improves detection accuracy .

• **Saving Processed Videos** : The face-cropped frames are then saved as new video files, forming a processed dataset that contains only facial video segments. This filtered data is used in the subsequent training and testing phases to ensure that the model focuses on relevant visual information.

Once the data is preprocessed, it is divided into training and testing sets to develop and evaluate the model. This section of the design includes:

• **Data Splitting** : The processed dataset is split into training and testing sets. This ensures that the model has both training data to learn from and testing data to validate its performance.

.• **Data Loader** : A data loader component is used to efficiently load the training videos and their corresponding labels (real or fake) into the model.

• **Deepfake Detection Model** : The core of the detection process leverages a dual-model architecture, utilizing both a Convolutional Neural Network (CNN) for feature extraction and a Recurrent Neural Network (RNN) for temporal analysis.

• **ResNext Feature Extraction (CNN)** : ResNext, a variant of CNN, is employed for extracting spatial features from the frames. By identifying subtle inconsistencies in pixel patterns, textures, and facial expressions, ResNext effectively highlights anomalies that may indicate deepfake content.

• **LSTMVideoClassification (RNN)** : Long Short-Term Memory (LSTM) networks, a type of RNN, are integrated to analyze temporal sequences. This model tracks frame-to-frame changes, providing insights into motion inconsistencies and unnatural transitions that are common in deepfakes.

• **Confusion Matrix** : A confusion matrix is generated to assess the model's classification performance on the test data. This metric provides insight into the accuracy, precision, and recall, offering a comprehensive view of the model's effectiveness in distinguishing real content from fake.

• **Export Trained Model** : Once validated, the trained model is exported for deployment, allowing it to be used for real-time or batch predictions on new video content.

• **Upload and Preprocess Video** : Users can upload a video, which then undergoes the same preprocessing steps—splitting into frames, face detection, and face cropping.

• **Load Trained Model** : The trained model is loaded to begin the prediction process on the uploaded video.

• **Real/Fake Classification** : Using the ResNext and LSTM components, the model performs a spatial and temporal analysis of the video, classifying it as either "Real" or "Fake." This output provides a clear, actionable result for the end user.

Working :

The proposed deepfake detection system operates through two main processes: the Training Flow and the Prediction Flow. Both processes utilize machine learning models to analyze both spatial and temporal features to detect manipulated videos accurately.

In the Training Flow, the system begins by preparing a labeled dataset of real and fake videos. These videos undergo preprocessing, where they are split into individual frames. Each frame is processed using face detection to isolate the face region, which is crucial for deepfake analysis. This allows the system to focus on the most relevant area for detection.

The face-cropped frames are then saved, creating a processed dataset that only contains the facial segments of each video. This dataset is divided into training and testing sets, enabling the model to be trained and evaluated effectively. The use of these distinct sets ensures that the model can be optimized without overfitting.

To capture key features, the system uses a dual-model architecture. A Convolutional Neural Network (CNN), specifically ResNext, is employed for spatial feature extraction. This model identifies subtle details such as pixel patterns, textures, and facial expressions, which are essential for distinguishing real videos from deepfakes.

In addition to spatial analysis, a Recurrent Neural Network (RNN), specifically Long Short-Term Memory (LSTM), is used to analyze the temporal features of the video. The LSTM model tracks changes across frames, detecting inconsistencies in motion and facial expressions that are typical in deepfake videos. This combination of spatial and temporal models helps the system detect both visual discrepancies and motion irregularities.

Once the model is trained, it is evaluated using a confusion matrix. This evaluation measures accuracy, precision, and recall to assess how well the model can differentiate real content from fake. If the model meets performance standards, it is exported for deployment in the prediction phase.

In the Prediction Flow, users can upload a video for real-time analysis. The video undergoes the same preprocessing steps as in the training phase, including frame splitting, face detection, and face cropping. The trained model is then used to analyze the video, with the ResNext model extracting spatial features and the LSTM component performing temporal analysis. Based on this combined analysis, the system classifies the video as either "Real" or "Fake," providing the result to the user.

4.

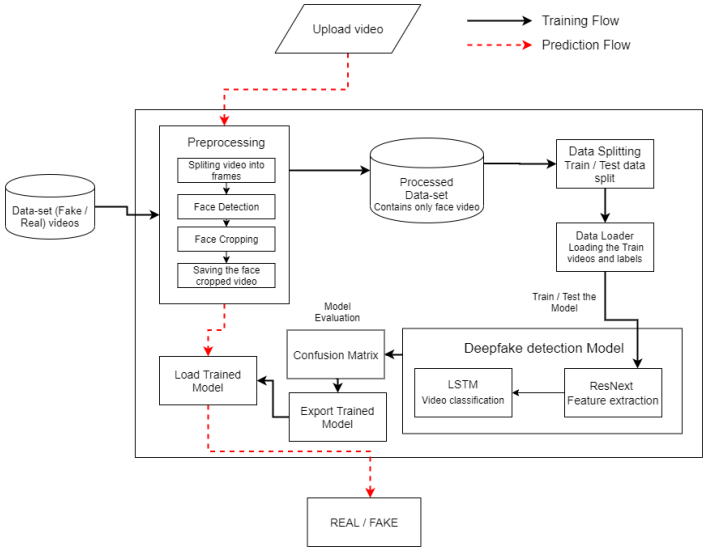

Fig 1: System Architecture

5. RESULTS AND DISCUSSION

The deepfake detection system demonstrated impressive performance in distinguishing between real and manipulated videos, utilizing the dual-model architecture of ResNext (CNN) for spatial feature extraction and LSTM (RNN) for temporal analysis. Upon evaluation, the system achieved high accuracy, precision, and recall on the testing dataset, indicating its effectiveness in identifying subtle visual discrepancies and motion irregularities present in deepfake content. The confusion matrix analysis further validated the model's capability to reduce false positives and false negatives, which is crucial for real-time deployment in applications like social media and video verification. The combination of CNN for pixel-level analysis and RNN for frame-to-frame sequence analysis ensured that both spatial and temporal features were adequately captured, contributing to robust deepfake detection across a variety of video manipulations.

While the system demonstrated strong performance, it also encountered challenges, especially with highly compressed videos or videos with low-resolution faces, where detecting deepfakes became more difficult. These limitations highlight the importance of continuously updating the training dataset to include such variations. Additionally, the system's reliance on face detection means that any errors in this preprocessing step can negatively impact overall accuracy. Despite these challenges, the proposed deepfake detection system provides a comprehensive and reliable approach, combining state-of-the-art techniques to offer an effective solution for real-time deepfake detection in various practical scenarios. Future improvements could focus on enhancing model robustness to compression artifacts and expanding its ability to detect more diverse forms of media manipulation beyond face-swapping.

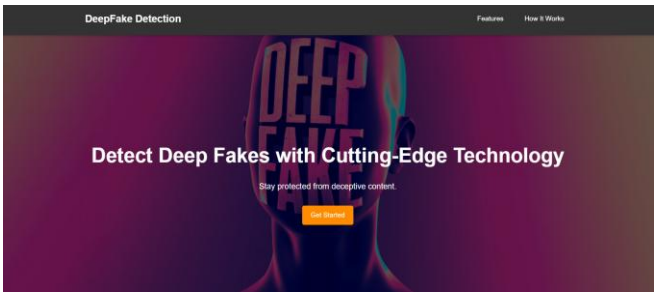

Fig 2: Home Page

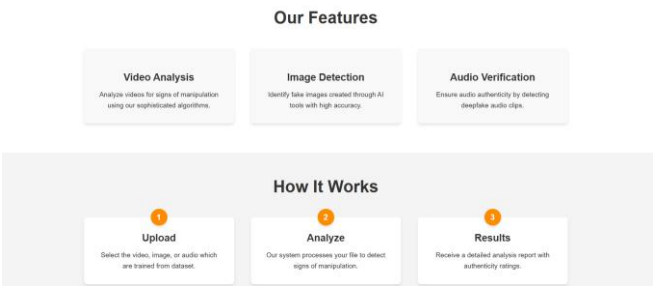

Fig 3: Home Page

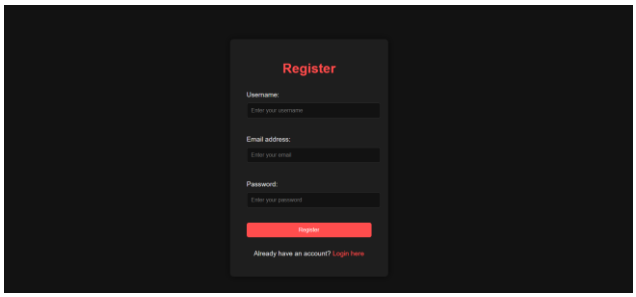

Fig 4: Register Page

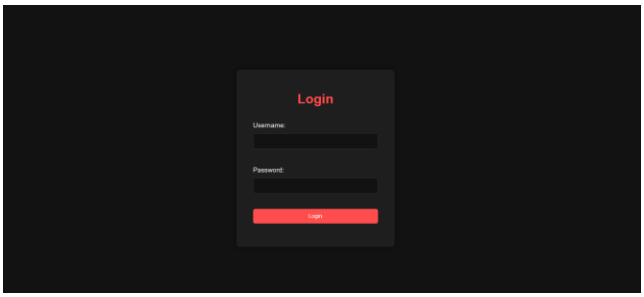

Fig 5: Login Page

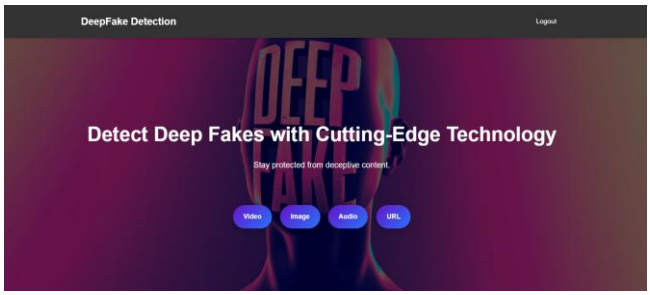

Fig 6: Detect Page

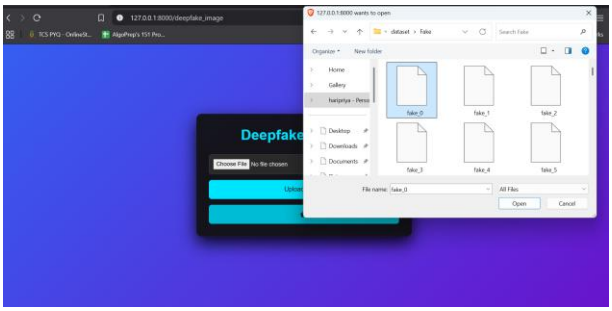

Fig 7: Upload Page

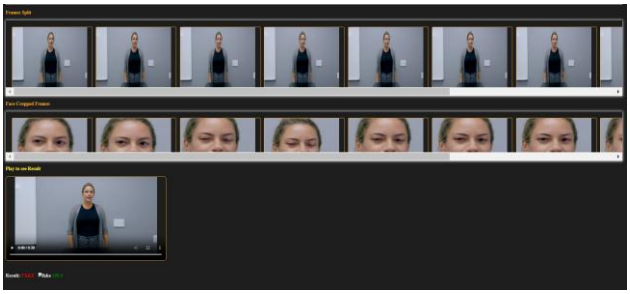

Fig 8: Video Results Page

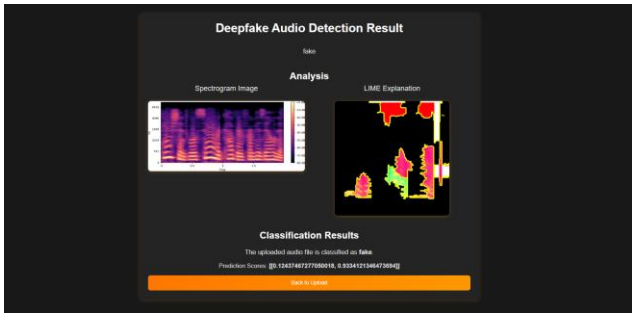

Fig 9: Audio Results Page

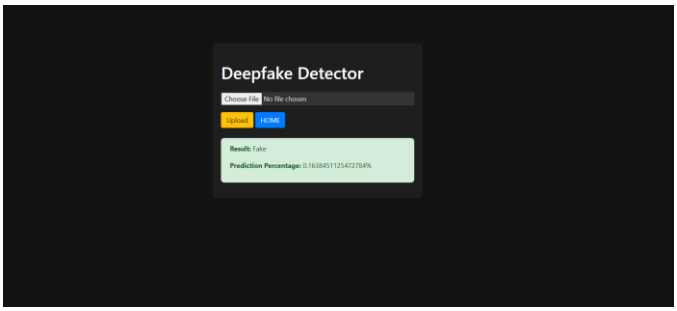

Fig 8: Summary Page

6. CONCLUSION

In conclusion, the deepfake detection system offers a comprehensive and reliable solution to combat the growing challenge of digital media manipulation. By utilizing Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs) for temporal analysis, the system excels at identifying subtle inconsistencies in both frame-level details and motion patterns. The CNNs effectively capture small distortions in texture, color, and facial movements, while the RNNs track changes over time to detect unnatural expressions and motions. This dual approach ensures high accuracy and makes the system an invaluable tool for sectors like journalism and law enforcement, where content authenticity is crucial for public trust and credibility.

The system's adaptability ensures it will remain effective as deepfake technology evolves. With rapid advancements in machine learning fueling the creation of increasingly sophisticated deepfakes, the system is designed to incorporate updated algorithms, refining its analysis to stay ahead of new manipulation techniques. This ongoing adaptability allows the system to serve as a proactive defense against emerging threats in the digital landscape, helping institutions, journalists, law enforcement, and users maintain the integrity of digital content. By preventing the spread of misinformation and ensuring the authenticity of media, this detection system is a vital tool in the effort to safeguard public trust in digital information.

REFERENCES

[1] DeepFake MNIST+: A DeepFake Facial Animation Dataset 023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) | 979-8-3503-4107-2/23/$31.00 ©2023 IEEE | DOI: 1 0.1109/CSDE59766.2023.10487745 .

[2] M. Westerlund, "A Survey on Deepfake Video Detection," Technology Innovation Management Review, vol. 9, pp. 39–52, Nov. 2021. DOI: 10.22215/timreview/1282.

[3] P. Korshunov and S. Marcel, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face ForgeryDetection," CoRR,vol. abs/1812.08685, 2020. arXiv: 1812 . 08685. [On line]. Available: http://arxiv.org/abs/1812.08685

[4] X. Xuan, B. Peng, J. Dong, and W. Wang, "DFT-MF: Enhanced Deepfake Detection Using Mouth Movement and Transfer Learning," CoRR, vol. abs/1902.11153, 2021. arXiv: 1902 . 11153. [On line]. Available: http://arxiv.org/abs/1902.11153.

[5] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Deep Insights of Deepfake Technology," CoRR,vol. abs/1611.09577, 2020. arXiv: 1611 . 09577. [On line]. Available: http://arxiv.org/abs/1611.09577.

[6] K. Olszewski, Z. Li, C. Yang, et al., "Deep Learning for Deepfakes Creation and Detection," in 2017 IEEEInternational Conference on Computer Vision (ICCV), 2022, pp. 5439–5448. DOI: 10 . 1109 / ICCV. 2017.580.

[7] I. Perov, D. Gao, N. Chervoniy, et al., "DeepFake Detection for Human Face Images and Videos," CoRR, vol. abs/2005.05535, 2021. arXiv: 2005.05535.[Online]. Available:https://arxiv.org/abs/2005.05535.

[8] K. Dale, K. Sunkavalli, M. Johnson, D. Vlasic, W. Matusik, and H. Pfister, "A GAN-Based Model of Deepfake Detection in Social Media," ACM Transactions on Graphics, vol. 30, pp. 1–10, Dec. 2022. DOI: 10.1145/2024156.2024164.